

Means and covariance functions for spatial compositional data: an axiomatic approach

Denis Allard^a, Thierry Marchant^b

^aBiostatistique et Processus Spatiaux (BioSP), INRA, Avignon, France.

^bGhent University, Ghent, Belgium.

Corresponding author:

D. Allard. INRA, UR546 BioSP, Site Agroparc

84914 Avignon, FRANCE

Tel: +33 432722171; Fax: +33 432722182

allard@avignon.inra.fr

Abstract

This work focuses on the characterization of the central tendency of a sample of compositional data. It provides new results about theoretical properties of means and covariance functions for compositional data, with an axiomatic perspective. Original results that shed new light on the geostatistical modeling of compositional data are presented. As a first result, it is shown that the weighted arithmetic mean is the only central tendency characteristic verifying a small set of axioms, namely reflexivity and marginal stability. Moreover, the weights must be identical for all components of the compositional vector. This result has deep consequences on the spatial multivariate covariance modeling of compositional data. In a geostatistical setting, it is shown as a second result that the proportional model of covariance functions (i.e. the product of a covariance matrix and a single correlation function) is the only model that provides identical kriging for all components of the compositional data. As a consequence of these two results, the proportional model of covariance function is the only covariance model compatible with reflexivity and marginal stability.

Keywords Aitchinson geometry; central tendency; functional equation; geostatistics

Authors are listed alphabetically. They contributed equally.

1 Introduction

This work focuses on the characterization of the central tendency of a sample of compositional data and on consequences regarding its geostatistical modeling. Compositional vectors are subject to a constant-sum constraint. Their modeling and their analysis is therefore very different from those of unconstrained multivariate vectors. They convey information about relative, not absolute, values of components. Formally, a compositional datum with p variables, $\mathbf{x} = (x^1, \dots, x^p)$, has positive components that add up to a constant, say κ . Without loss of generality, one can set $\kappa = 1$ for the rest of this paper. Compositional data thus belong to the positive simplex of dimension $p - 1$

$$\mathbb{S}^{p-1} = \{(x^1, \dots, x^p) : x^k \geq 0 \text{ for } k = 1, \dots, p, \text{ and } x^1 + \dots + x^p = 1\}. \quad (1)$$

The central tendency of a sample of compositional data, which will also be a compositional data, denoted $\mathbf{M}(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{S}^{p-1}$. Aitchinson (1989) states that arithmetic means are “clearly useless as a measure of location because it falls outside of the array of compositions and is indeed very atypical of the data set” and that the normalized geometric mean (i.e. the back-transform of the arithmetic mean of the log-ratios) “serves equally well for curved data sets and for more linear and elliptical data sets”. On ternary sub-compositions of hongite, Sharp (2006) noted that in many instances the geometric means also falls outside of the array of compositions, a fact already pointed out in Shurtz (2000). Probably the most striking example is the Hongite 2 artificial data sets, where all samples are projected on a line parallel to one side of the triangle. The arithmetic average is shown to belong to the same line, while the geometric average does not (Sharp, 2006, Fig. 2).

As an alternative to arithmetic or geometric means, Sharp (2006) proposed the graph median. It is built from the minimum spanning tree, which is the graph connecting all points of the data set whose total length (sum of the length of the edges of the graph) is minimum. The distance used in the simplex is the half-taxi metric (Miller, 2002) $d(\mathbf{x}, \mathbf{x}') = 0.5 \sum_{k=1}^p |x^k - x'^k|$. The graph median is then obtained by iteratively pruning the outermost branches of the tree until only one point or a pair of points remain. This point or the mid-point between the pair of points is then the graph median. The minimum spanning tree is not always unique, in which case a tie breaking rule is necessary. Very often, the graph median is one of the sample points of the data. Otherwise it is the mid-point between two close data samples. By construction it is located in the innermost region of the data-set, thus defining a central tendency (Sharp, 2006). It is however complicated to compute and it cannot be easily related to the estimation of a total quantity or indeed to most statistical or geostatistical analysis. Moreover, it is not continuous with respect to the data values. This alternative will not be considered in this work.

The statistical analysis of compositional data has received great attention in the last three

decades. Compositional data are often transformed into a new vector of log-ratios of the components of the vector (Aitchinson, 1982, 1986; Pawlowsky-Glahn and Olea, 2004; Egozcue, 2003). These transformations provide one-to-one mappings onto a real space, thereby allowing usual multivariate statistics methods to be applied to the transformed data. Any statement made in the transformed space being easily translated back into the compositional space. Billheimer (2001) and Pawlowsky-Glahn and Egozcue (2002) showed that the simplex \mathbb{S}^{p-1} equipped with the scalar product

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{p} \sum_{i=1}^p \sum_{j=i+1}^p \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \sum_{i=1}^p \ln \frac{x_i}{g(\mathbf{x})} \ln \frac{y_i}{g(\mathbf{y})}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{S}^{p-1}, \quad (2)$$

where $g(\mathbf{x}) = [x^1 x^2 \dots x^p]^{1/p}$ is the geometric mean of the components of \mathbf{x} , induces a distance and thus a geometry on the simplex, called the Aitchinson geometry. An orthonormal basis within the simplex was proposed in (Egozcue, 2003; Mateu-Figueras, 2011) where the data (x^1, \dots, x^p) are transformed by means of the isometric log-ratio (ilr) transformations

$$u^i = [i(i+1)]^{-1/2} \ln \frac{x^1 x^2 \dots x^i}{(x^{i+1})^i}, \quad i = 1, \dots, p-1, \quad (3)$$

with $\mathbf{u} = (u^1, \dots, u^{p-1}) \in \mathbb{R}^{p-1}$. Some properties of compositional data and their parameters can be worked out directly within the Aitchinson geometry of the simplex. However, when backtransforming these to the standard Euclidean space of the raw data, unexpected behaviors can take place. Following (Mateu-Figueras, 2011, pp. 35–36), let us consider the following example, with $\mathbf{x} = (0.6, 0.3, 0.1)$ and $\mathbf{x}' = (0.3, 0.3, 0.4)$ being two compositions in \mathbb{S}^2 . Their ilr transforms $\mathbf{u} = (0.490, 1.180)$ and $\mathbf{u}' = (0.000, -0.235)$ correspond to orthonormal coordinates with respect to the ilr transformation. We can then apply standard operations on these coordinates. For example, their arithmetic mean is $\bar{\mathbf{u}} = (0.245, 0.4725)$. Once backtransformed by the inverse operations of Eq. (3), we obtain

$$\text{ilr}^{-1}(\bar{\mathbf{u}}) = \tilde{\mathbf{x}} = (0.459, 0.325, 0.216),$$

which is nothing but the closure to one of the geometric mean of \mathbf{x} and \mathbf{x}' . An unexpected and intriguing fact is that even though the second coordinates are equal to 0.3 for both data, the second coordinate of $\tilde{\mathbf{x}}$ is increased by 8.3%.

In the simplex, whenever one component is increased (resp. decreased), some or all other components must decrease (resp. increase) in order to verify the sum constraint, a fact that has consequences on the conditions one wishes to impose on \mathbf{M} . Consider a dataset $\mathbf{x}_1, \dots, \mathbf{x}_n$ of compositional data with q variables (i.e., $\mathbf{x}_i = (x_i^1, \dots, x_i^q)$) to be grouped into $p < q$ variables in two different ways, thereby defining two new datasets $\mathbf{y}_1, \dots, \mathbf{y}_n$ and $\mathbf{z}_1, \dots, \mathbf{z}_n$. Consider further that the first variable is identical in the two groupings, i.e., $y_i^1 = z_i^1$, $i = 1, \dots, n$. For instance, with $q = 4$ and $p = 3$, one grouping is $1, \{2, 3\}, 4$ and

the other one is 1, 2, {3, 4}. One does not expect the mean of the first variable to depend on the grouping of the other variables. This is the case for the arithmetic mean. Indeed, for the first grouping, the k th component of the arithmetic mean is $M_a^k(\mathbf{y}_1, \dots, \mathbf{y}_n) = n^{-1} \sum_{i=1}^n y_i^k$ and a similar expression holds for $M_a^k(\mathbf{z}_1, \dots, \mathbf{z}_n)$. Since the first variable is common to both datasets, it is clear that $M_a^1(\mathbf{y}_1, \dots, \mathbf{y}_n) = M_a^1(\mathbf{z}_1, \dots, \mathbf{z}_n)$. The arithmetic mean is said to verify marginal stability.

Let us now compute the normalized geometric mean independently for each variable. Then, for the first grouping

$$M_g^k(\mathbf{y}_1, \dots, \mathbf{y}_n) = \frac{(\prod_{i=1}^n y_i^k)^{1/n}}{\sum_{l=1}^p (\prod_{i=1}^n y_i^l)^{1/n}},$$

for $k = 1, \dots, p$, with a similar expression for the second grouping. Since all variables are involved in the computation of each component of the normalized geometric mean, equality of the mean of the first variable in the two groupings will be different, i.e., $M_g^1(\mathbf{y}_1, \dots, \mathbf{y}_n) \neq M_g^1(\mathbf{z}_1, \dots, \mathbf{z}_n)$. In other words, the normalized geometric mean does not verify marginal stability. A similar result is obtained for any log-ratio transform. Marginal stability is an important property that one may choose to impose or not when analyzing compositional data. It plays a critical role in our results.

For applications, such as oil and mining industry or soil depollution, where compositions correspond to fractions of soil type or rock type, the above examples are puzzling. Based on samples of compositional data, practitioners working in these fields need to compute unbiased estimates of the total amount of material for a given volume, i.e., unbiased estimates of the mean of the fraction. Unanswered questions arise: Under which conditions is marginal stability always verified on central tendencies? What are the conditions for having unbiased estimates of means? In a geostatistical setting, what are the multivariate covariance functions compatible with marginal stability?

With an axiomatic perspective, this paper provides new mathematical results about theoretical properties of means and covariance functions for compositional data, that shed new light on their geostatistical modeling. We start from a set of axioms that correspond to what we feel are quite natural conditions: reflexivity (which will be defined later) and marginal stability. Other sets of axioms would lead to different mathematical properties. For example, subcompositional dominance is violated by the Euclidean distance between compositional data (Egozcue, 2011). It is not the purpose of the present work to discuss the relevance of sets of axioms for the analysis of data. We merely derive “if and only if” relationships between one given set of axioms and some mathematical properties, which we believe are useful in some situations.

As a first result, it will be shown that the weighted arithmetic mean is the only central tendency characteristic verifying reflexivity and marginal stability. Moreover, the weights

must be identical for all components of the compositional vector. This first result is independent of any geometry and thus holds in the standard Euclidean space of the raw data and in the Aitchinson geometry of the simplex. It has deep consequences on the multivariate covariance modeling. It is well known (and easy to verify) that if the multivariate covariance function belongs to the family of proportional models (i.e. it is the product of a covariance matrix and a single correlation function), the kriging weights are identical for all variables. It will be shown that the converse also holds for all multivariate random fields, compositional or not. As a second result on compositional data, it is thus established that the proportional model of covariance functions is, in the Euclidean space of raw data, the only covariance model verifying reflexivity and marginal stability. This second result does not necessarily hold within the Aitchinson geometry of the simplex.

The paper is structured as follows. Section 2 provides a quick presentation of the axiomatic characterization of the possible definitions of means. Section 3 presents our first main result on the mean of compositional data. In Sect. 4, the consequences of this theorem on the covariance functions are shown. These results are then discussed in Sect. 5.

2 A primer to axiomatic definitions of means

2.1 Means for univariate data

There exists many ways of summarizing a sample x_1, \dots, x_n of a variable defined on an interval $E \subset \mathbb{R}$ into a single value, usually called the mean, and sometimes also called central tendency. The most widely used are the arithmetic, geometric or harmonic means, the mode, the median, and more generally any desired quantile. They all verify different properties. For example, the arithmetic mean $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ is the real number m minimizing $\sum_{i=1}^n (m - x_i)^2$. It is unique, easy to compute, but sensitive to large values or outliers. The median is the real number m minimizing $\sum_{i=1}^n |m - x_i|$. It is sometimes not unique, but robust to large values and outliers. An axiomatic characterization of the arithmetic mean was obtained as early as Kolmogorov (1930). Let us denote $M(x_1, \dots, x_n)$ a mapping from $E^n \rightarrow \mathbb{R}$, which will be the central tendency, or the mean, of the sample and let us impose the following, quite natural, conditions, called axioms. Kolmogorov's characterization follows.

- K1 *Continuity and strict monotonicity:* $M(x_1, \dots, x_n)$ is continuous and strictly monotonic in each of its arguments. Strict monotonicity states that if one of the values x_i increases, the central tendency must also increase. Continuity imposes that small variations in one of the values x_i leads to small variations of M .
- K2 *Symmetry:* $M(x_1, \dots, x_n)$ is symmetric, i.e., it is identical for any permutation of the sample.

K3 *Reflexivity*: The central tendency of identical values is equal to their common value, i.e., $M(x, \dots, x) = x$.

K4 *Associativity*: A subset of the sample can be replaced by its central tendency with no effect on the total central tendency:

$$M(x_1, \dots, x_m, x_{m+1}, \dots, x_n) = M(x_*, \dots, x_*, x_{m+1}, \dots, x_n)$$

where $x_* = M(x_1, \dots, x_m)$.

Theorem 1 (Kolmogorov, 1930) *Conditions K1 to K4 hold if and only if the central tendency M has the form*

$$M(x_1, \dots, x_n) = \phi^{-1} \left(\frac{\phi(x_1) + \dots + \phi(x_n)}{n} \right), \quad (4)$$

where ϕ is a continuous strictly increasing function on E , called the generating function.

Functions M having the form of Eq. (4) are called quasi-arithmetic means in the functional equation literature (Aczél and Dhombres, 1989; Matkowski, 2010). It is worth noting that neither the mode nor the median belong to this family. The median is not continuous, while neither the mode nor the median verify K4. Quasi-arithmetic cover a wide range of well known means: if $\phi(x) = x$, M is the arithmetic mean. On $E = (0, +\infty)$, $\phi(x) = \ln x$ leads to the geometric mean, while $\phi(x) = x^{-1}$ leads to the harmonic mean. More generally, when $\phi(x) = x^\alpha$, $\alpha \neq 0$ and $x \in (0, +\infty)$ the associated mean is called the power mean.

2.2 Means for multivariate data

For multivariate data, each data is a vector of length p , i.e. $\mathbf{x}_i = (x_i^1, \dots, x_i^p)$, with $i = 1, \dots, n$. In this case, the central tendency is a vector $\mathbf{M} = (M^1, \dots, M^p)$, where M^k is a mapping from $E^{p \times n} \rightarrow \mathbb{R}$. Possibilities are numerous. A tempting simplification is to impose that each component M^k depends only on the values of the corresponding variable. In this case one builds one central tendency per variable (e.g., the arithmetic mean or the median), independently of all other variables. When applying Kolmogorov's Theorem to each variable independently, the associated central tendency M^k is a quasi-arithmetic mean if and only if conditions K1-K4 hold. When $E \subset \mathbb{R}$, the multivariate arithmetic mean is the point $\mathbf{M}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \mathbf{m} \in \mathbb{R}^p$ minimizing $\sum_{i=1}^n \|\mathbf{m} - \mathbf{x}_i\|_p^2$, where $\|\cdot\|_p$ is the Euclidean distance in \mathbb{R}^p . In this case, each component of $\mathbf{M}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ is the arithmetic mean of the corresponding variable.

Although being natural and probably widely used, the componentwise simplification is by no means the only mathematical possibility. For instance, in analogy to the definition of the median for univariate data, the multivariate median is the vector \mathbf{m} minimizing

$\sum_{i=1}^n \|\mathbf{m} - \mathbf{x}_i\|_p$. It is unique whenever the points are not colinear (Vardi and Zhang, 2000). Its k -th coordinate, m^k , depends on the values of all coordinates of the samples $\mathbf{x}_1, \dots, \mathbf{x}_n$. Many alternative multivariate medians can be defined, such as those based on the notion of statistical depth (Liu et al., 1999; Zuo and Sefling, 2000) or the graph median proposed in Sharp (2006).

3 An axiomatic characterization of the mean

As seen in the Introduction, there exists many possible approaches for defining a mean for compositional data, and they all verify different properties. Inspired by the axiomatic approach briefly summarized in Sect. 2, a new characterization theorem for compositional data is presented. It relies on functional equation arguments and does not necessitate the use of inner products and distances. It is independent of any geometry on the simplex.

Consider a sample of fixed size n of compositional data belonging to the positive simplex \mathbb{S}^{p-1} , corresponding to a regionalized variable $\mathbf{x}(\cdot)$ defined in \mathbb{R}^d , sampled at sites $\mathbf{s}_1, \dots, \mathbf{s}_n$. For the sake of simplicity, we will write $\mathbf{x}_i = \mathbf{x}(\mathbf{s}_i)$. We want to characterize the mappings $\mathbf{M} = (M^1, \dots, M^p)$ which associate to each vector $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ a vector of means $\mathbf{M}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ with components $M^k(\mathbf{x}_1, \dots, \mathbf{x}_n)$, for $k = 1, \dots, p$. Notice that at this stage $M^k(\mathbf{x}_1, \dots, \mathbf{x}_n)$ can depend on $\mathbf{x}_1, \dots, \mathbf{x}_n$ and not just on x_1^k, \dots, x_n^k so that the definition of \mathbf{M} is very general.

Conditions imposed for the characterization of the central tendency of spatial compositional data differ to those imposed for Kolmogorov's characterization. Among Kolmogorov's conditions, only one is absolutely necessary: reflexivity. As discussed in the previous section, marginal stability is also imposed. Our main theorem, characterizing the means for compositional data, follows.

C1 Reflexivity: for all $\mathbf{x} \in \mathbb{S}^{p-1}$, $\mathbf{M}(\underbrace{\mathbf{x}, \dots, \mathbf{x}}_n) = \mathbf{x}$.

C2 Marginal stability: for any $k = 1, \dots, p$, any $i = 1, \dots, n$ and any vectors $\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}'_i$ in \mathbb{S}^{p-1} , if $x_i^k = x'^k_i$, then

$$M^k(\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n) = M^k(\mathbf{x}_1, \dots, \mathbf{x}'_i, \dots, \mathbf{x}_n).$$

In presence of spatial correlation, spatial symmetry is in general not desired, but can be an option. For the sake of completeness it is recalled.

C3 Symmetry:

$$M^k(\mathbf{x}_1, \dots, \mathbf{x}_n) = M^k(\mathbf{x}_{\sigma(1)}, \dots, \mathbf{x}_{\sigma(n)})$$

for any permutation σ of $\{1, \dots, n\}$ and any $\mathbf{x}_1, \dots, \mathbf{x}_n$ of \mathbb{S}^{p-1} .

Theorem 2 (Characterization of the mean) Consider compositional data $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ with $p \geq 3$ and $n \geq 2$. The mapping \mathbf{M} satisfies conditions C1-C2 if and only if, for $k \in \{1, \dots, p\}$,

$$M^k(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \lambda_i x_i^k, \quad (5)$$

for some real numbers $(\lambda_1, \dots, \lambda_n)$ verifying $\sum_{i=1}^n \lambda_i = 1$. Furthermore, if symmetry holds, then $\lambda_i = 1/n$ for all $i \in \{1, \dots, n\}$.

The key ingredients of the proof are marginal stability and closure to one. Marginal stability imposes that for each $k = 1, \dots, p$, M^k depends only on x_1^k, \dots, x_n^k . Then, closure to one of the mean implies that the generating function ϕ must be the identity function, i.e. that the only means are linear combinations. Note that the same result holds if data sum to another constant than 1. Note also that when working with compositional data, which are positive and bounded, continuity and monotonicity conditions are not strictly needed to establish Theorem 2. A formal proof is presented in Appendix.

Note also that this result holds whenever there are at least three components. With two components there is only one independent variable and hence many more means verifying axioms C1-C2. As an example, it can easily be shown that the Kolmogorov means characterized in Eq. (4)

$$M^k(\mathbf{x}_1, \dots, \mathbf{x}_n) = \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(x_i^k) \right)$$

also verify marginal stability and closure to one whenever $\phi : [0, 1] \rightarrow \mathbb{R}$ is such that $\phi(t) = 1 - \phi(1 - t)$, with $\phi(0) = 0$ and $\phi(1) = 1$. Indeed,

$$\begin{aligned} M^1(\mathbf{x}_1, \dots, \mathbf{x}_n) + M^2(\mathbf{x}_1, \dots, \mathbf{x}_n) &= \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(x_i^1) \right) + \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(1 - x_i^1) \right) \\ &= \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(x_i^1) \right) + \phi^{-1} \left(\sum_{i=1}^n \lambda_i (1 - \phi(x_i^1)) \right) \\ &= \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(x_i^1) \right) + \phi^{-1} \left(1 - \sum_{i=1}^n \lambda_i \phi(x_i^1) \right) \end{aligned}$$

Since $\phi(t) = 1 - \phi(1 - t)$ implies $\phi^{-1}(u) = 1 - \phi^{-1}(1 - u)$ for $0 \leq u \leq 1$, the last equality implies

$$M^1(\mathbf{x}_1, \dots, \mathbf{x}_n) + M^2(\mathbf{x}_1, \dots, \mathbf{x}_n) = \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(x_i^1) \right) + 1 - \phi^{-1} \left(\sum_{i=1}^n \lambda_i \phi(x_i^1) \right) = 1.$$

The case $p = 2$ is not of central interest for the analysis of compositional data. In the rest of this work, it is thus assumed that $p \geq 3$.

4 Consequences in geostatistics

Within a statistical setting, the condition $\sum_{i=1}^n \lambda_i = 1$ in Eq. (5) corresponds to an unbiasedness condition. In this setting, Theorem 2 states that means of compositional data satisfying axioms C1-C2 correspond to linear unbiased estimators. This Theorem does not provide any criterion for choosing the weights in Eq. (5). In particular, it does not make explicit reference to the location of the data.

In geostatistics, it is well known that, provided the covariance function is known, the best linear predictor of the vector of the mean is the kriging of the mean (Cressie, 1993; Chilès and Delfiner, 2012). As an immediate corollary of Theorem 2, we obtain the following two facts

1. Axioms C1-C2 are verified if and only if the mean \mathbf{M} is an unbiased linear combination of the data $\mathbf{x}_1, \dots, \mathbf{x}_n$.
2. kriging of the mean is thus the best (minimum variance) predictor of the mean verifying conditions C1-C2.

It must also be noticed that the weights $(\lambda_1, \dots, \lambda_n)$ do not depend on the variable for which the mean is computed. In other words, the weights are identical for all variables. This fact has deep consequences on the covariance modeling of spatial compositional data, which are now examined in details. The usual geostatistical setting is considered. The compositional data $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ is a sample of a second order stationary multivariate random field $\mathcal{X}(\mathbf{s})$ with $\mathbf{s} \in \mathbb{R}^d$. For the simplicity of exposition, consider $p \geq 3$ and $n \geq 2$. Under the assumption of second order stationarity, the multivariate covariance model of $\mathcal{X}(\mathbf{s})$ is defined by a matrix of functions

$$\begin{pmatrix} C_{11}(\mathbf{h}) & \cdots & C_{1p}(\mathbf{h}) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{h}) & \cdots & C_{pp}(\mathbf{h}) \end{pmatrix}, \mathbf{h} \in \mathbb{R}^d, \quad (6)$$

which must be positive definite (Chilès and Delfiner, 2012). For a given set of n locations $(\mathbf{s}_1, \dots, \mathbf{s}_n)$, this model induces a multivariate covariance $np \times np$ block-matrix of the form

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{p1} & \cdots & \mathbf{C}_{pp} \end{pmatrix}, \quad (7)$$

where each $n \times n$ matrix \mathbf{C}_{kl} is such that its elements are $[\mathbf{C}_{kl}]_{ij} = C_{kl}(\mathbf{s}_j - \mathbf{s}_i)$.

Kriging weights depend on the multivariate covariance model and on sample locations. It is well known (and easy to verify) that if the multivariate covariance function belongs to the

family of proportional models (i.e. the product of a covariance matrix and a single correlation function), the kriging weights are identical for all variables.

The following Theorem shows that the converse also holds. It is stated independently of the compositional data setting considered so far. A formal link to the compositional data setting will be made later.

Theorem 3 *Let us consider a second-order stationarity multivariate random field of dimension $p \geq 2$. For all $n \geq 2$ and all sample design $(\mathbf{s}_1, \dots, \mathbf{s}_n)$, the co-kriging weights of the p means are equal for all variables if and only if the multivariate covariance model is a proportional model, i.e. for $1 \leq k, l \leq p$*

$$C_{kl}(\mathbf{h}) = \sigma_{kl}\rho(\mathbf{h}), \quad \mathbf{h} \in \mathbb{R}^d,$$

for some $p \times p$ covariance matrix $\mathbf{\Sigma} = [\sigma_{kl}]_1^p$ and some correlation function $\rho(\mathbf{h})$.

The proof of this Theorem is deferred to the Appendix. Particular cases of this model are absence of correlation between variables, with $\sigma_{kl} = 0$ for all $k \neq l$. This particular case must be ruled out when considering compositional data since closure to one imposes negative correlation between some variables. Theorems 2 and 3 considered together imply that, when kriging the means of second order stationary compositional data, the only multivariate covariance function model verifying conditions C1-C2 is the proportional model

$$\mathbf{C}(\mathbf{h}) = \mathbf{\Sigma} \otimes \rho(\mathbf{h}),$$

where $\rho(\mathbf{h})$ is a correlation function and $\mathbf{\Sigma}$ a valid covariance matrix for compositional data. This is now formally stated in the following Theorem, in which $\text{tr}(\mathbf{A})$ denotes the trace operator of the square matrix \mathbf{A} .

Theorem 4 *Let $\mathcal{X}(\mathbf{s})$ be a multivariate compositional random field as described above and let \mathbf{M} be a multivariate mean of this data. Then, the following propositions are equivalent:*

1. $\mathbf{M} = (M^1, \dots, M^p)$ verifies C1-C2 and is such that $\text{tr}\{\text{Var}(\mathbf{M})\}$ is minimum;
2. For each variable $k = 1, \dots, p$, M^k is a linear combination of (x_1^k, \dots, x_n^k) , where the weights are independent of the variable k . The weights $\lambda_1, \dots, \lambda_n$ are solution of a unique kriging system.
3. The multivariate covariance function model is a proportional model.

Proof. The proof consists in collecting the results established in Theorems 2 and 3.

- a) According to Theorem 2, “ \mathbf{M} verifies C1-C2” is equivalent to “ M^k is a linear combination of (x_1^k, \dots, x_n^k) only. Moreover, the weights are identical for all variables $k = 1, \dots, p$ ”. Therefore, imposing $\text{tr}\{\text{Var}(\mathbf{M})\}$ to be minimum is equivalent to imposing $\text{Var}(M^k)$ to be minimum for each $k = 1, \dots, p$, i.e. it is equivalent to impose M^k to be equal to the kriging of the mean of (x_1^k, \dots, x_n^k) , for some unique kriging system. Hence $1 \Leftrightarrow 2$.
- b) In theorem 3 it is proved that $2 \Leftrightarrow 3$.

In conclusion, the three statements are equivalent. \square

5 Discussion

In this work several original results are established:

- Firstly, it is shown that for compositional data the only means that verify simultaneously reflexivity and marginal stability are weighted arithmetic means, to which kriging belongs. The simultaneous requirement of marginal stability and closure to one is the main reason leading to this result.
- Secondly, a very general result in multivariate geostatistics is established. It is shown that the only multivariate covariance model for which the kriging weights are identical for all components is the proportional model. To the best of our knowledge, this result has not been shown earlier. It has consequences for the modeling of compositional data, but it also has an interest on its own.
- Thirdly, the combination of these two results is that, within a geostatistical setting in the Euclidean space of the raw data, the only covariance model leading to a kriging that verifies simultaneously reflexivity and marginal stability is the proportional model.

In summary, when performing the statistical analysis of compositional data in the standard Euclidean geometry of real space it is impossible, at the same time, to verify axioms C1-C2 and to consider a complex modeling using log-ratio transformations, complex covariance models, or both. It is a kind of impossibility result, in the spirit of Arrow’s impossibility theorem (Arrow, 1950). This results might be perceived as a disappointing one. Depending on the problem considered, one has either to relax marginal stability or to restrict the modeling to the very simple proportional covariance models and kriging of the mean on raw data.

Axioms are sets of properties we choose, and this choice can be discussed. In Egozcue and Pawłowski-Glahn (2011), a principle of coherence is chosen to rule out the Euclidean distance between vectors of proportions. In essence, this is also an axiomatic approach,

somehow similar to the one followed here. Is the “principle of coherence” a better condition than marginal stability (C2)? This is an interesting debate and an open question that would necessitate to first provide an operative definition of “better”. This work was intended to provide new mathematical results, and to open a new research direction on compositional data, based on an axiomatic approach. It must be considered as complementary to the usual model-based approach, based on log-ratios. Instead of starting from a model and exploring its properties, we start from a set of properties (the axioms) and derive the class of models verifying this set of axioms. We leave to further work to explore if there is an “if and only if” relationship between the “principle of coherence” and the use of log-transforms. Usual geostatistical tools such as covariances, variograms and kriging can be redefined in this Hilbert space (Pawlowsky-Glahn and Egozcue, 2002; Tolosana-Delgado et al., 2011). An interesting development would be to reformulate an axiomatic approach in the simplex \mathbb{S}^{p-1} equipped with the isometric log-ratio transform (Egozcue, 2003) and check whether a result similar to Theorem 2 holds.

References

- Arrow KL (1950) A Difficulty in the Concept of Social Welfare. *J Polit Econ* 58(4):328–346.
- Aczél J (1966) Lectures on functional equations and their applications. Academic Press.
- Aczél J and Dhombres J (1989) Functional equations in several variables. Cambridge University Press
- Aitchinson J (1982) The statistical analysis of compositional data. *J Royal Stat Soc Ser B (Stat Methodol)* 44(2):139–177
- Aitchinson J. (1986) The statistical analysis of compositional data. Chapman and Hall
- Aitchinson J (1989) Measures of location of compositional data sets. *Math Geol* 24(4):365–379
- Billheimer D, Guttorp P and Fagan WF (2001) Statistical Interpretation of Species Composition. *J Am Stat Assoc* 96:1205–1214
- Cressie N (1993) Statistics for Spatial Data, Revised Edition. Wiley
- Chilès JP and Delfiner P (2012) Geostatistics: modeling spatial uncertainty; Second Edition. John Wiley & Sons
- Egozcue JJ, Pawlowksy-Glahn V, Figureas GM and Barceló-Vidal C (2003) Isometric Logratio Transformations for Compositional Data Analysis. *Math Geol* 35(3):279–300

- Egozcue JJ and Pawlowksy-Glahn V (2011) Basic Concepts and Procedures. In *Compositional Data Analysis: Theory and Applications*, Eds. Pawlowsky-Glahn V, Buccianti A. John Wiley & Sons
- Kolmogorov A (1930) Sur la notion de la moyenne. *Atti R Accad Naz Lincei Mem Cl Sci Fis Mat Natur Sez* 12:323–343
- Liu RY, Parelius JM and Singh K (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *Ann Stat* 27(3):783–858
- Mateu-Figueras G, Pawlowksy-Glahn V and Egozcue JJ (2011) The principle of working on coordinates. In *Compositional Data Analysis: Theory and Applications*, Eds. Pawlowsky-Glahn V, Buccianti A. John Wiley & Sons
- Matkowski J (2010) Generalized weighted quasi-arithmetic means. *Aequ Math* 79:203–212
- Miller WE (2002) Revisiting the Geometry of a Ternary Diagram With the Half-Taxi Metric. *Math Geol* 34(3):275–290
- Pawlowsky-Glahn V and Olea RA (2004) *Geostatistical Analysis of Compositional Data*. Oxford University Press
- Pawlowsky-Glahn V and Egozcue JJ (2002) BLU Estimators and Compositional Data. *Math Geol* 34(3):259–274
- Sharp WE (2006) The graph median – a stable alternative measure of central tendency for compositional data sets. *Math Geol*, 38:221–229
- Shurtz RF (2000) Comment on “Logratios and natural laws in compositional data analysis” by J Aitchison. *Math Geol* 32:645–647
- Tolosana-Delgado R, van den Boogaart K and Pawlowsky-Glahn V (2011) Geostatistics for compositions. In *Compositional Data Analysis: Theory and Applications*, Eds. Pawlowsky-Glahn V, Buccianti A. John Wiley & Sons
- Vardi Y, Zhang CH (2000) The multivariate L_1 -median and associated data depth. *Proc Natl Acad of Sc* 97(4):1423–1426
- Zuo Y, Serfling R (2000). General notions of statistical depth function. *Ann stat* 28(2): 461–482

Appendix A: Proof of Theorem 2

Our definition of the simplex is a closed set, i.e. some compositions are allowed to be equal to 0. While this might be a problem for the definition of log-ratios, it will not be a problem for us.

1/ Necessity of the conditions. It is simple to check that (5) verifies all conditions C1-C2.

2/ Sufficiency of the conditions. By marginal stability, we have

$$M^k(\mathbf{x}_1, \dots, \mathbf{x}_n) = F_k(x_1^k, \dots, x_n^k)$$

for some function $F_k : [0, 1]^n \rightarrow [0, 1]$. Choose any $l, l', l'' \in \{1, \dots, p\}$ and suppose $x_i^k = 0$ for all $k \neq l, l', l''$ and all $i \in \{1, \dots, n\}$. Since $\sum_{k=1}^p M^k(\mathbf{x}_1, \dots, \mathbf{x}_n) = 1$ and $x_i^{l''} = 1 - x_i^l - x_i^{l'}$ for all $i \in \{1, \dots, n\}$, it is the case that

$$F_l(x_1^l, \dots, x_n^l) + F_{l'}(x_1^{l'}, \dots, x_n^{l'}) + F_{l''}(1 - x_1^l - x_1^{l'}, \dots, 1 - x_n^l - x_n^{l'}) = 1. \quad (8)$$

Let us define the mapping $G : [0, 1]^n \rightarrow [0, 1]$ by $G(u_1, \dots, u_n) = 1 - F_{l''}(1 - u_1, \dots, 1 - u_n)$. Equation (8) then becomes

$$F_l(x_1^l, \dots, x_n^l) + F_{l'}(x_1^{l'}, \dots, x_n^{l'}) = G(x_1^l + x_1^{l'}, \dots, x_n^l + x_n^{l'}), \quad (9)$$

for all $x_1^l, x_1^{l'} \in [0, 1]$ such that $x_1^l + x_1^{l'} \leq 1$. In particular, it holds for all $x_1^l, x_1^{l'} \in [0, 1/2]$. Equation (9) is a generalized Pexider equation and, because F_l , $F_{l'}$ and G are bounded, its unique solution is

$$\begin{aligned} F_l(u_1, \dots, u_n) &= \lambda_1 u_1 + \dots + \lambda_n u_n + \gamma_l, & \forall u_1, \dots, u_n \in [0, 1/2], \\ F_{l'}(u_1, \dots, u_n) &= \lambda_1 u_1 + \dots + \lambda_n u_n + \gamma_{l'}, & \forall u_1, \dots, u_n \in [0, 1/2], \\ G(u_1, \dots, u_n) &= \lambda_1 u_1 + \dots + \lambda_n u_n + \gamma_l + \gamma_{l'}, & \forall u_1, \dots, u_n \in [0, 1], \end{aligned}$$

for some real numbers $\lambda_1, \dots, \lambda_n, \gamma_l$ and $\gamma_{l'}$ (Aczél, 1966, p.302). The expression for G yields $F_{l''}(u_1, \dots, u_n) = 1 - G(1 - u_1, \dots, 1 - u_n) = \lambda_1 u_1 + \dots + \lambda_n u_n + \beta_{l''}$ for all $u_1, \dots, u_n \in [0, 1]$ and for some real $\beta_{l''}$.

Since our choice of components l, l' and l'' in the above reasoning is arbitrary, we obtain $M^k(\mathbf{x}_1, \dots, \mathbf{x}_n) = F_k(x_1^k, \dots, x_n^k) = \sum_{i=1}^n \lambda_i x_i^k + \beta_k$, for all $k = 1, \dots, p$ and all x_1^k, \dots, x_n^k in $[0, 1]$. By reflexivity, $F_k(u, \dots, u) = \sum_{i=1}^n \lambda_i u + \beta_k = u$, for all $u \in [0, 1]$ and for all $k = 1, \dots, p$. This is possible only if $\beta_k = 0$ for all $k = 1, \dots, p$ and $\sum_{i=1}^n \lambda_i = 1$. This concludes the proof. \square

Appendix B: Proof of Theorem 3

In the following, $\mathbf{1}_n$ denotes a vector of ones of length n , \mathbf{I}_n denotes the identity matrix of dimension n and $\mathbf{0}_{p,q}$ denotes a $p \times q$ matrix of zeros. If \mathbf{A} is a $m \times n$ matrix and \mathbf{B} is a $p \times q$ matrix, the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is the $mp \times nq$ block matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}.$$

In particular, the matrix $\mathbf{J} = \mathbf{I}_p \otimes \mathbf{1}_n$ is the $np \times p$ matrix

$$\begin{pmatrix} \mathbf{1}_n & \cdots & \mathbf{0}_{n,1} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{n,1} & \cdots & \mathbf{1}_n \end{pmatrix}.$$

For each variable $k = 1, \dots, p$, the kriging of the mean, \hat{m}_k , is a linear combination of the data

$$\hat{m}_k = \sum_{l=1}^p \mathbf{X}_l^\top \boldsymbol{\lambda}_l^k,$$

where $\mathbf{X}_l = (x_{1,l}, \dots, x_{n,l})^\top$ and $\boldsymbol{\lambda}_l^k = (\lambda_{1,l}^k, \dots, \lambda_{n,l}^k)^\top$. Unbiasedness conditions impose

$$\mathbf{1}_n^\top \boldsymbol{\lambda}_k^k = 1 \quad \text{and} \quad \mathbf{1}_n^\top \boldsymbol{\lambda}_l^k = 0, \quad \text{for } l \neq k.$$

Let $\boldsymbol{\lambda}^k$ be the stacked np -vector $\boldsymbol{\lambda}^k = ((\boldsymbol{\lambda}_1^k)^\top, \dots, (\boldsymbol{\lambda}_p^k)^\top)^\top$ and let $\mathbf{\Lambda} = (\boldsymbol{\lambda}^1, \dots, \boldsymbol{\lambda}^p)$ be the $(np \times p)$ matrix of kriging weights. When solved simultaneously, the p kriging equations are, in matrix notation,

$$\begin{pmatrix} \mathbf{C} & \mathbf{J} \\ \mathbf{J}^\top & \mathbf{0}_{p,p} \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{np,p} \\ \mathbf{I}_p \end{pmatrix}, \quad (10)$$

where $\boldsymbol{\mu}$ is the $p \times p$ matrix of Lagrange multipliers. The matrix \mathbf{C} arises from a valid model of covariance functions. If one excludes multiple values at the same location, it is invertible. The general solution for $\mathbf{\Lambda}$ therefore verifies

$$\mathbf{C}\mathbf{\Lambda} = \mathbf{J}(\mathbf{J}^\top \mathbf{C}^{-1} \mathbf{J})^{-1}. \quad (11)$$

For each $k = 1, \dots, p$, we wish to impose that $\boldsymbol{\lambda}^k$ is a vector of zeros except at coordinates corresponding to the k -th variable where the weights are equal to a common vector $\boldsymbol{\lambda}_0$, i.e. $\boldsymbol{\lambda}^k = (\mathbf{0}_{1,n(k-1)}, \boldsymbol{\lambda}_0^\top, \dots, \mathbf{0}_{1,n(p-k)})^\top$. Hence, $\mathbf{\Lambda} = \mathbf{I}_p \otimes \boldsymbol{\lambda}_0$. Thus, Eq. (11) becomes

$$\mathbf{C}(\mathbf{I}_p \otimes \boldsymbol{\lambda}_0) = \mathbf{J}(\mathbf{J}^\top \mathbf{C}^{-1} \mathbf{J})^{-1}. \quad (12)$$

In order to have a closer look at this condition, let us denote $\mathbf{D} = (\mathbf{J}^t \mathbf{C}^{-1} \mathbf{J})^{-1}$, which is a $p \times p$ matrix with elements D_{kl} . Then, Eq. (12) is equivalent to

$$\mathbf{C}_{kl} \boldsymbol{\lambda}_0 = D_{kl} \mathbf{1}_n, \quad \forall 1 \leq k, l \leq p. \quad (13)$$

Since \mathbf{C} is invertible, \mathbf{C}_{kk} is invertible and $D_{kk} \neq 0$, for all $k = 1, \dots, p$. Hence, plugging $D_{kk} \mathbf{C}_{kk}^{-1} \mathbf{1}_n = \boldsymbol{\lambda}_0$ into Eq. (13) leads to

$$\frac{D_{kk}}{D_{ll}} \mathbf{C}_{kk}^{-1} \mathbf{C}_{ll} = \mathbf{I}_n, \quad \forall 1 \leq k, l \leq p.$$

This condition shows that $\mathbf{C}_{kk} = \sigma_{kk} \mathbf{R}$, where \mathbf{R} is a correlation matrix. With a similar argument, one can show that $\mathbf{C}_{kl} = \sigma_{kl} \mathbf{R}$ when $k \neq l$. In conclusion, there is a single correlation matrix for describing all direct and cross covariance matrices,

$$\mathbf{C} = \boldsymbol{\Sigma} \otimes \mathbf{R}.$$

In other words, the model is proportional. □